Lecture 16: Inference for Two Samples

STAT 515 — Statistical Methods I

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Outline

1. From one sample to two: a motivating trial
2. Large-sample inference for $\mu_1 - \mu_2$ (two-sample z)
3. Small-sample inference: the pooled t
4. When the variances differ: Welch's t
5. Paired differences
6. Summary: which two-sample procedure, when

Outline

1. From one sample to two: a motivating trial
2. Large-sample inference for $\mu_1 - \mu_2$ (two-sample z)
3. Small-sample inference: the pooled t
4. When the variances differ: Welch's t
5. Paired differences
6. Summary: which two-sample procedure, when

Where We Are

Every test so far compared *one* sample against a **number**: $\mu = \mu_0$, $p = p_0$, $\sigma^2 = \sigma_0^2$.

Today

Lots of real studies compare **two groups** — treatment vs. placebo, machine 1 vs. machine 2, before vs. after. The parameter of interest becomes a **difference**, and the question becomes "is that difference zero?"

- The **six-step recipe does not change**. Only the estimate and its standard error change.
- The one genuinely new idea is that **how the two samples were collected** — independently, or in pairs — decides which procedure is correct.

A Motivating Study: The EPREDA Trial

Uzan et al., *The Lancet* (1991)

A randomized, placebo-controlled trial asking whether low-dose aspirin (with or without dipyridamole) prevents fetal growth retardation. Pregnant women were randomized to **placebo** or **treatment**, and birth weight was recorded.

Group	n	Mean (g)	SD (g)
Placebo	73	2526	848
Treated	156	2751	670

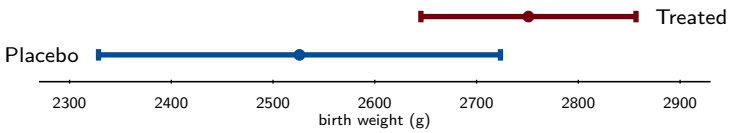
The authors report: "*Mean birth weight was significantly higher in the treated than in the placebo group . . . difference* 225 *g*, $p = 0.029$."

A Tempting — and Wrong — Shortcut

We already know how to build a one-sample CI. Do it for each group:

$$\text{Placebo:}\quad 2526 \pm 1.99\,\frac{848}{\sqrt{73}} = (\,2328.5,\ 2723.5\,)$$

$$\text{Treated:}\quad 2751 \pm 1.97\,\frac{670}{\sqrt{156}} = (\,2645.3,\ 2856.7\,)$$



The two intervals **overlap**, which tempts the conclusion "no significant difference." **That reasoning is not valid.**

Why Overlapping Intervals Prove Nothing

- Each interval answers a question about *one* mean. Neither one is an interval for the **difference**.
- The standard error of the difference is **not** the sum of the two individual margins of error — variances add, **margins of error do not**.
- Two intervals can overlap and the difference can still be significantly different from zero (as here); they can also fail to overlap when the difference is not significant.

The rule

To answer a question about a difference, build a confidence interval for the difference and test a hypothesis about the difference.

Outline

1. From one sample to two: a motivating trial
2. Large-sample inference for $\mu_1 - \mu_2$ (two-sample z)
3. Small-sample inference: the pooled t
4. When the variances differ: Welch's t
5. Paired differences
6. Summary: which two-sample procedure, when

Notation: Two Populations, Two Samples

	Population 1	**Population 2**
Mean, standard deviation	$\mu_1,\ \sigma_1$	$\mu_2,\ \sigma_2$
Sample size	n_1	n_2
Observations	$X_{11}, \ldots, X_{1n_1}$	$X_{21}, \ldots, X_{2n_2}$
Sample mean	$\overline{X_1}$	$\overline{X_2}$
Sample variance	s_1^2	s_2^2

- The parameter of interest is $\mu_1 - \mu_2$; the natural estimate is $\overline{X_1} - \overline{X_2}$.
- **Decide which population is "1" before you start** and keep it fixed — the sign of every answer depends on it.

The Sampling Distribution of $\overline{X_1} - \overline{X_2}$

For two **independent** random samples:

Mean and variance of the difference

$$E(\overline{X_1} - \overline{X_2}) = \mu_1 - \mu_2 \qquad \sigma^2_{\overline{X_1} - \overline{X_2}} = \text{Var}(\overline{X_1}) + \text{Var}(\overline{X_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- $\overline{X_1} - \overline{X_2}$ is an **unbiased** estimator of $\mu_1 - \mu_2$.
- **Variances add even though we subtract the means** — the two samples are independent, so their uncertainties accumulate.
- Hence $\sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Why the Variances Add

Two rules from earlier in the course do all the work:

(i) $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$ for any constant a.

(ii) $X,\ Y$ **independent** $\implies \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Putting them together

$$\begin{aligned}
\mathrm{Var}\left(\overline{X_1} - \overline{X_2}\right) &= \mathrm{Var}\left(\overline{X_1} + (-1)\overline{X_2}\right) \\
&= \mathrm{Var}(\overline{X_1}) + (-1)^2\,\mathrm{Var}(\overline{X_2}) && \text{by (ii), then (i)} \\
&= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} && \text{since } \mathrm{Var}(\overline{X}) = \sigma^2/n
\end{aligned}$$

The step that surprises people: $(-1)^2 = +1$, so the minus sign **squares away** — subtracting the means does **not** subtract the uncertainty.

Two Consequences Worth Remembering

- **Standard errors never add.** If $SE(\overline{X_1}) = 3$ and $SE(\overline{X_2}) = 4$, then

$$SE(\overline{X_1} - \overline{X_2}) = \sqrt{3^2 + 4^2} = \mathbf{5} \qquad \text{not } 7, \text{ and not } 1.$$

- **Step (ii) needs independence.** In general $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X, Y)$, and the covariance term vanishes **only** when X and Y are independent.

Looking ahead

That covariance term is not a nuisance — it is an **opportunity**. When two measurements are taken on the *same* subject they are positively correlated, $\mathrm{Cov}(X, Y) > 0$, and the variance of the difference **shrinks**. That is exactly why **pairing** works, in Section 5.

Fact

For two independent large samples ($n_1 \geq 30$ and $n_2 \geq 30$), the Central Limit Theorem gives

$$\overline{X_1} - \overline{X_2} \dot{\sim} N\left(\mu_1 - \mu_2,\ \sigma^2_{\overline{X_1}-\overline{X_2}}\right), \qquad \sigma^2_{\overline{X_1}-\overline{X_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

In practice σ_1 and σ_2 are unknown, and with large samples we simply substitute the sample standard deviations:

$$\sigma_{\overline{X_1}-\overline{X_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Exactly the one-sample story, one layer up: a statistic, its standard error, and a normal reference curve.

Confidence Interval for $\mu_1 - \mu_2$ (Large Samples)

$100(1-\alpha)\%$ confidence interval

$$(\overline{X_1} - \overline{X_2}) \pm z_{\alpha/2}\,\sigma_{\overline{X_1}-\overline{X_2}} \;\approx\; (\overline{X_1} - \overline{X_2}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Interval		Interpretation
$(-, -)$	$\Longrightarrow$	$\mu_1 - \mu_2 < 0$, i.e. $\mu_1 < \mu_2$
$(+, +)$	$\Longrightarrow$	$\mu_1 - \mu_2 > 0$, i.e. $\mu_1 > \mu_2$
$(-, +)$	$\Longrightarrow$	0 is plausible; no evidence that $\mu_1 \neq \mu_2$

Reading the signs is the whole interpretation. An interval that covers 0 is exactly an interval that fails to establish a difference.

Hypothesis Test for $\mu_1 - \mu_2$ (Large Samples)

Two-sample z test

$$H_0: \ \mu_1 - \mu_2 = D_0 \qquad Z = \frac{(\overline{X_1} - \overline{X_2}) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \ \dot{\sim} \ N(0,1) \ \text{ under } H_0$$

Alternative	Rejection region
$H_a: \ \mu_1 - \mu_2 > D_0$	$\{ Z > z_\alpha \}$
$H_a: \ \mu_1 - \mu_2 < D_0$	$\{ Z < -z_\alpha \}$
$H_a: \ \mu_1 - \mu_2 \neq D_0$	$\{ \lvert Z \rvert > z_{\alpha/2} \}$

Almost always $D_0 = 0$: the hypothesis of "no difference." Then $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$.

Conditions for the Large-Sample (z) Procedures

Assumptions for inference about $\mu_1 - \mu_2$

- The two samples are **randomly selected** from the two target populations, **in an independent manner**.
- The sample sizes are both large: $n_1 \geq 30$ and $n_2 \geq 30$.

- **No normality assumption on either population** — the CLT takes care of that, which is exactly what large samples buy us.
- **Independence is not optional.** If the two samples are matched or measured on the same subjects, the formula for $\sigma_{\overline{X_1} - \overline{X_2}}$ above is simply wrong. We return to that case in Section 5.

Example 1: Repair Times for Two Kinds of Equipment

Example 1

An investigation of repair times for two kinds of photocopying equipment produced the following results.

Type	# of repairs	Mean	Std. dev.
1	$n_1 = 60$	$\overline{x_1} = 84.2$	$s_1 = 19.4$
2	$n_2 = 60$	$\overline{x_2} = 91.6$	$s_2 = 18.8$

(a) At the .01 level, do repair times appear to differ? Also calculate the p-value.

(b) Construct a 99% confidence interval for the mean difference and interpret it.

(c) What assumptions are needed?

$\mu_1,\ \mu_2 =$ the mean repair times of type 1 and type 2 equipment.

Example 1(a): Test Statistic

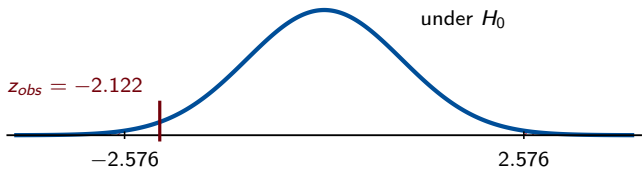
$$H_0: \ \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0 \qquad \text{vs.} \qquad H_a: \ \mu_1 \neq \mu_2 \iff \mu_1 - \mu_2 \neq 0$$

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{84.2 - 91.6}{\sqrt{\frac{19.4^2}{60} + \frac{18.8^2}{60}}} = \frac{-7.4}{3.488} = \textbf{-2.122}$$

Both samples are large ($n_1 = n_2 = 60 \geq 30$), so the two-sample z procedure applies.

Rejection region: $\{\, |Z| > z_{\alpha/2} \,\} = \{\, |Z| > z_{.005} \,\} = \{\, |Z| > 2.576 \,\}$

Example 1(a): Decision and p-Value



Since $|-2.122| \not> 2.576$, the statistic is **not** in the rejection region: **fail to reject** H_0 at the .01 level.

$$p\text{-value} = 2\,P_{H_0}(Z \leq -2.122) = 2(0.017) = \mathbf{0.034} \qquad 0.034 \not< 0.01$$

Conclusion

At the .01 significance level we do **not** have sufficient evidence to say that the two mean repair times are different.

Example 1(b)–(c): Interval and Assumptions

$$(\overline{x_1} - \overline{x_2}) \pm z_{.005}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (84.2 - 91.6) \pm 2.576\sqrt{\frac{19.4^2}{60} + \frac{18.8^2}{60}}$$
$$= -7.4 \pm 8.984 = (\,-16.384,\ 1.584\,)$$

1. We are 99% confident that the true mean difference in repair times is between -16.384 and 1.584 minutes.
2. The interval is of the form $(-, +)$: it **covers** 0, so there is no significant difference between the two mean repair times — **the same conclusion the test reached**.

(c) Assumptions

Two **independent random samples**, each of size at least 30.

Both groups are large, so the same z procedure applies with 1 = treated, 2 = placebo:

$$\text{SE} = \sqrt{\frac{670^2}{156} + \frac{848^2}{73}} = 112.8, \qquad z_{obs} = \frac{2751 - 2526}{112.8} = \frac{225}{112.8} = \mathbf{1.99}$$

$$95\%\ \text{CI}: \quad 225 \pm 1.96(112.8) = 225 \pm 221.1 = (\,3.9,\ 446.1\,)$$

The interval is $(+, +)$ and excludes 0; $p \approx 0.046$. We **reject** H_0: mean birth weight is higher under treatment.

A question worth asking

The paper reports the interval [129, 321], much narrower than ours. **Why?** Reported summaries are not always enough to reproduce a published analysis — a habit of checking is worth acquiring.

Outline

1. From one sample to two: a motivating trial
2. Large-sample inference for $\mu_1 - \mu_2$ (two-sample z)
3. Small-sample inference: the pooled t
4. When the variances differ: Welch's t
5. Paired differences
6. Summary: which two-sample procedure, when

Small Samples: What Must Be Assumed

When n_1 or n_2 is small the CLT no longer rescues us, so we must assume more about the populations.

Fact

For two independent samples, *provided*

- both populations are **normally distributed**, and
- the two population variances are **equal**, $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

$$\frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t\,(df = n_1 + n_2 - 2)$$

The Pooled Sample Variance

Pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then $\text{Var}(\overline{X_1} - \overline{X_2}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, so there is **one** variance to estimate, not two.
- s_p^2 is a **weighted average** of s_1^2 and s_2^2, with weights equal to the degrees of freedom $n_1 - 1$ and $n_2 - 1$: the larger sample gets more say.
- The two samples contribute $n_1 - 1$ and $n_2 - 1$ degrees of freedom, hence $df = n_1 + n_2 - 2$.

Confidence Interval and Test (Pooled t)

$100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$

$$(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \qquad df = n_1 + n_2 - 2$$

Test of $H_0: \ \mu_1 - \mu_2 = D_0$

$$t = \frac{(\overline{X_1} - \overline{X_2}) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad df = n_1 + n_2 - 2$$

Rejection regions: $\{t > t_{n_1+n_2-2,\alpha}\}$, $\{t < -t_{n_1+n_2-2,\alpha}\}$, or $\{|t| > t_{n_1+n_2-2,\alpha/2}\}$ — the same three shapes as always.

Conditions for the Small-Sample (t) Procedures

Assumptions for inference about $\mu_1 - \mu_2$

- The two samples are **randomly selected** from the two target populations, **in an independent manner**.
- Both sampled populations are **approximately normal**.
- The population variances are **equal**: $\sigma_1^2 = \sigma_2^2$.

How seriously to take them

The pooled t is **robust** to moderate non-normality, especially when $n_1 = n_2$. Equal variance matters most when the sample sizes are **very unequal**; a rule of thumb is to worry once the larger s exceeds about twice the smaller.

Example 2: Test Scores at Two High Schools

Example 2

The following are scores on a standardized test for two independent random samples of students from two high schools.

School A: $78, 84, 81, 78, 76, 83, 79, 75, 85, 81$
School B: $85, 75, 83, 87, 80, 79, 88, 94, 87, 82$

(a) At the .05 level, does it appear that school A has lower scores than school B? Also calculate the p-value. (b) Construct a 95% confidence interval for the difference between mean scores. (c) What assumptions are needed?

$\mu_A =$ the mean score of school A; $\mu_B =$ the mean score of school B.

Example 2: Summary Statistics and Pooled Variance

$$n_A = 10, \quad \overline{x_A} = 80, \quad s_A = 3.367 \qquad\qquad n_B = 10, \quad \overline{x_B} = 84, \quad s_B = 5.396$$

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{9(3.367)^2 + 9(5.396)^2}{18} = \mathbf{20.227}$$

$$\text{SE} = \sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} = \sqrt{20.227\left(\tfrac{1}{10} + \tfrac{1}{10}\right)} = \sqrt{4.045} = 2.011, \qquad df = 18$$

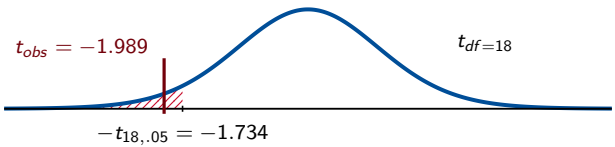
s_B is about 1.6 times s_A — within the range where pooling is reasonable, and $n_A = n_B$, which makes the procedure especially forgiving.

Example 2(a): Test Statistic and Rejection Region

$$H_0: \ \mu_A - \mu_B = 0 \ (\mu_A = \mu_B) \qquad \text{vs.} \qquad H_a: \ \mu_A - \mu_B < 0 \ (\mu_A < \mu_B)$$

$$t = \frac{(\overline{x_A} - \overline{x_B}) - 0}{\sqrt{s_p^2 \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{80 - 84}{2.011} = \frac{-4}{2.011} = \textbf{-1.989}$$

Rejection region: $\{ t < -t_{18,.05} \} = \{ t < -1.734 \}$



Example 2(a): Decision and Conclusion

- Since $-1.989 < -1.734$, the statistic **falls in the rejection region**: **reject** H_0 at the .05 level.
- p-value $= P_{H_0}(t_{18} < -1.989) =$ **0.031** $< .05$ — the same decision.

Conclusion

At the .05 level we have sufficient evidence that the mean score at school A is **significantly lower** than the mean score at school B.

Note the p-value is *one*-tailed here, because H_a points in one direction. Had we asked only whether the schools *differ*, the p-value would be $2(0.031) = 0.062$ and we would **not** reject.

Example 2(b)–(c): Interval and Assumptions

$$(\overline{x_A} - \overline{x_B}) \pm t_{18,\,.025}\sqrt{s_p^2\left(\tfrac{1}{n_A} + \tfrac{1}{n_B}\right)} = -4 \pm 2.101(2.011) = (\,-8.225,\ 0.225\,)$$

Why doesn't this interval match the test?

The two-sided 95% interval covers 0, yet the *one*-sided test rejected. There is no contradiction: **a one-sided test pairs with a one-sided interval.** The matching 95% upper bound is

$$(\overline{x_A} - \overline{x_B}) + t_{18,\,.05}\,(2.011) = -4 + 1.734(2.011) = -0.513$$

so we are 95% confident the true mean difference lies **below zero**.

(c) Assumptions: two independent random samples; both populations approximately normal; $\sigma_A^2 = \sigma_B^2$.

Outline

1. From one sample to two: a motivating trial
2. Large-sample inference for $\mu_1 - \mu_2$ (two-sample z)
3. Small-sample inference: the pooled t
4. When the variances differ: Welch's t
5. Paired differences
6. Summary: which two-sample procedure, when

Relaxing the Equal-Variance Assumption

- The two-sample t test is a **powerful** tool *when its assumptions are satisfied*.
- When the **sample sizes are equal** ($n_1 = n_2$) the equal-variance assumption can simply be relaxed: use $\hat{\sigma}_{\overline{X_1}-\overline{X_2}} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$.
- Even when the sample sizes are **not** equal and the variances are **not** equal, an **approximate** interval or test is obtained by modifying the degrees of freedom.

Terminology

This approximation goes by several names: **Welch's t**, the **separate-variance t**, or the **Satterthwaite approximation**. It is R's `t.test()` default.

Welch's t and the Satterthwaite Degrees of Freedom

Test statistic and approximate reference distribution

$$t = \frac{(\overline{X_1} - \overline{X_2}) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \;\dot{\sim}\; t_{df} \qquad \text{with} \qquad df = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- The df is generally **not an integer**; round down to be conservative, or let software use it directly.
- It always lies between $\min(n_1 - 1,\, n_2 - 1)$ and $n_1 + n_2 - 2$: never more information than pooling, often a little less.
- The interval uses the same SE with $t_{df,\,\alpha/2}$.

Example 2 Revisited: Does Pooling Matter Here?

Procedure	**SE**	t	df
Pooled t (equal variances)	2.011	-1.989	18
Welch's t (separate variances)	2.011	-1.989	15.09

- With $n_A = n_B$ the two standard errors are **algebraically identical**, so the test statistic is unchanged.
- Only the degrees of freedom shrink, from 18 to about 15, which moves the one-sided p-value from 0.031 to 0.033.
- **Same conclusion either way** — which is the usual outcome with equal sample sizes and comparable spreads.

When sample sizes are badly unbalanced *and* the spreads differ, the two procedures can disagree; then prefer Welch.

Outline

1. From one sample to two: a motivating trial
2. Large-sample inference for $\mu_1 - \mu_2$ (two-sample z)
3. Small-sample inference: the pooled t
4. When the variances differ: Welch's t
5. Paired differences
6. Summary: which two-sample procedure, when

When the Two Samples Are Not Independent

Paired difference experiment

An experiment in which the observations come in **pairs** — the same subject measured twice, twins, matched patients, left and right eye — and the **differences within pairs** are analyzed.

- Also called a **pre–post design**, **repeated measures**, or a **matched-pairs design**.
- The two samples are **not independent**, so none of the formulas in Sections 2–4 apply.
- The fix is beautifully simple: reduce each pair to its **difference** d_i, and the problem becomes a **one-sample** problem about μ_d — which we already know how to solve.

Why Pairing Helps

- In many cases a paired difference experiment provides **more information** about the difference between population means than an independent samples experiment can.
- The effect of pairing is to **remove the larger source of variation** that would be present if two independent samples were chosen.
- The differencing removes sources of variation that tend to **inflate the variance of the difference** — subject-to-subject variability cancels out.

The trade-off

Pairing costs degrees of freedom: $n - 1$ for n pairs, instead of $n_1 + n_2 - 2$. **It is worth it whenever the pairing is effective** — that is, whenever the paired measurements are strongly related.

The Paired t Procedure

Let $d_i = X_{1i} - X_{2i}$ be the difference for the ith of n pairs, and $\mu_d = \mu_1 - \mu_2$ the mean difference.

Test and interval for μ_d

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}, \quad s_d^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}, \quad s_{\overline{d}} = \frac{s_d}{\sqrt{n}}$$

$$t = \frac{\overline{d} - \mu_{d_0}}{s_d/\sqrt{n}} \sim t_{n-1}, \qquad \text{CI: } \overline{d} \pm t_{n-1,\,\alpha/2}\,\frac{s_d}{\sqrt{n}}$$

Assumptions

The n **differences** are a random sample from an **approximately normal** population. **Nothing is assumed about the two original populations separately.**

Example 3: Biofeedback and Blood Pressure

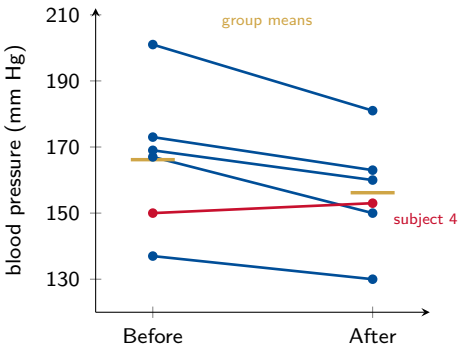
Example 3

An experiment was conducted to examine whether blood-pressure levels can be consciously lowered by persons trained in biofeedback exercises. Data were collected on six subjects trained in these exercises. Test at the .05 level of significance.

Subject	1	2	3	4	5	6
Before	137	201	167	150	173	169
After	130	181	150	153	163	160
d_i = Before − After	7	20	17	−3	10	9

Each subject is measured **twice** — the samples are paired, not independent.

Example 3: Each Subject Is Its Own Control



- Each line is **one subject** — the data are six *lines*, not twelve points.
- **Five of the six go down.** Only **subject 4** rises, by 3 mm Hg: the one negative d_i.
- The lines span 137–201 vertically, yet run roughly **parallel**. That spread is **between-subject** variation — noise to an independent-samples analysis, invisible to the paired test.

Reading a paired plot: **slopes carry the information**, heights do not.

Example 3: Hypotheses and Test Statistic

μ_B = mean blood pressure before training, μ_A = mean after; $\mu_d = \mu_B - \mu_A$.

$$H_0: \ \mu_B = \mu_A \iff \mu_d = 0 \qquad \text{vs.} \qquad H_a: \ \mu_B > \mu_A \iff \mu_d > 0$$

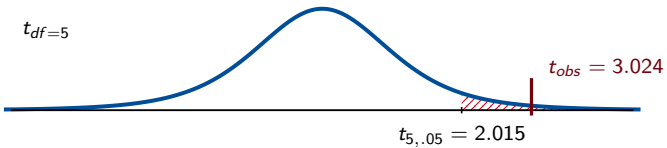
$$\overline{d} = \frac{7 + 20 + 17 - 3 + 10 + 9}{6} = \mathbf{10}, \qquad s_d = \mathbf{8.099}, \qquad s_{\overline{d}} = \frac{8.099}{\sqrt{6}} = 3.307$$

$$t = \frac{\overline{d} - 0}{s_d/\sqrt{n}} = \frac{10}{3.307} = \mathbf{3.024}, \qquad df = n - 1 = 5$$

"Consciously *lowered*" makes this an upper-tailed test on $\mu_d = \mu_B - \mu_A$.

Example 3: Rejection Region and Decision

Rejection region: $\{\, t > t_{5,.05} \,\} = \{\, t > 2.015 \,\}$



Since $3.024 > 2.015$ we **reject** H_0 at $\alpha = .05$; p-value $= P(t_5 > 3.024) =$ **0.015**.

Conclusion

At the .05 level we have sufficient evidence that biofeedback training **lowers** mean blood pressure. A 95% CI for μ_d is $10 \pm 2.571(3.307) = (\,1.50,\ 18.50\,)$, entirely above 0.

The Same Data, Analyzed Wrongly

Suppose we had ignored the pairing and run a two-sample pooled t test:

$$s_p = 19.47, \qquad t = \frac{(\overline{x_B} - \overline{x_A}) - 0}{\sqrt{s_p^2\left(\frac{1}{6} + \frac{1}{6}\right)}} = \frac{166.17 - 156.17}{11.24} = \mathbf{0.890}, \qquad df = 10$$

$$p\text{-value} = P(t_{10} > 0.890) = \mathbf{0.197} \quad \Longrightarrow \quad \text{fail to reject } H_0$$

Why the two answers differ so much

Blood pressure varies enormously *between* subjects (137 to 201). The independent-samples analysis buries the treatment effect inside that between-subject noise. **Differencing removes it**: the six differences range only from -3 to 20, and the signal becomes visible.

Using the wrong procedure here costs you the finding entirely.

Independent or Paired? How to Tell

Independent samples	**Paired samples**
Two separate groups of subjects	One group measured twice, or matched pairs
n_1 need not equal n_2	Always the same number in each column
Rows of the two groups have no link	Each row is *one* pair
Analyze $\overline{X_1} - \overline{X_2}$	Analyze the differences d_i
$df = n_1 + n_2 - 2$ (or Welch)	$df = n - 1$ pairs

The test to apply

Ask: can I shuffle one column without losing information? If yes, the samples are independent. If reordering one column breaks a meaningful correspondence — subject, twin, before/after — the data are **paired**.

Outline

1. From one sample to two: a motivating trial
2. Large-sample inference for $\mu_1 - \mu_2$ (two-sample z)
3. Small-sample inference: the pooled t
4. When the variances differ: Welch's t
5. Paired differences
6. Summary: which two-sample procedure, when

Choosing the Right Two-Sample Procedure

Design	Conditions	Statistic	Reference
Independent	$n_1, n_2 \geq 30$	$\dfrac{(\overline{x_1} - \overline{x_2}) - D_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$N(0,1)$
Independent	small n, normal, $\sigma_1^2 = \sigma_2^2$	$\dfrac{(\overline{x_1} - \overline{x_2}) - D_0}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}$	$t_{n_1+n_2-2}$
Independent	small n, normal, $\sigma_1^2 \neq \sigma_2^2$	$\dfrac{(\overline{x_1} - \overline{x_2}) - D_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$t_{\text{Satterthwaite}}$
Paired	differences approx. normal	$\dfrac{\overline{d} - \mu_{d_0}}{s_d/\sqrt{n}}$	t_{n-1}

Two questions, in order

(1) Are the samples independent or paired? (2) If independent: are the samples large, and are the spreads comparable?

Common Mistakes to Avoid

- **Comparing two one-sample confidence intervals** and concluding from the overlap. Build an interval for the **difference**.
- **Treating paired data as independent** (or the reverse). The design — not the data — decides the procedure.
- **Adding margins of error** instead of variances: $\sqrt{a} + \sqrt{b} \neq \sqrt{a+b}$.
- **Losing track of which group is "1."** The sign of $\overline{x_1} - \overline{x_2}$, of the interval, and of H_a must all agree.
- **Forgetting** $df = n_1 + n_2 - 2$ for the pooled t and $df = n - 1$ for the paired t.
- **Pairing a one-sided test with a two-sided interval** and calling the mismatch a contradiction.

Summary

- The parameter is $\mu_1 - \mu_2$; the estimate is $\overline{X_1} - \overline{X_2}$; for independent samples the variances **add**.
- **Large samples:** a z procedure with $\sqrt{s_1^2/n_1 + s_2^2/n_2}$; no normality needed.
- **Small samples, equal variances:** pool into s_p^2 and refer to $t_{n_1+n_2-2}$.
- **Small samples, unequal variances:** the separate-variance SE with Satterthwaite degrees of freedom.
- **Paired data:** reduce to the differences d_i and run a **one-sample** t test on μ_d with $df = n - 1$.
- A confidence interval for the difference that covers 0 says the same thing as a two-sided test that fails to reject.

Next: comparing two proportions, and comparing two variances.